

METHOD AND APPARATUS FOR EFFICIENT IDENTIFICATION OF
DUPLICATE AND NEAR-DUPLICATE DOCUMENTS AND TEXT SPANS
USING HIGH-DISCRIMINABILITY TEXT FRAGMENTS

ABSTRACT OF THE DISCLOSURE

5 Disclosed is a computer-assisted method for finding duplicate or near-
duplicate documents or text spans within a document collection by using high-
discriminability text fragments. Distinctive features of the documents or text spans are
identified. For each pair of documents or text spans with at least one distinctive feature
in common, the distinctive features of each document or text span are compared to
10 determine whether the pair is duplicates or near-duplicates. An apparatus for performing
this computer-assisted method is also disclosed.